几乎必然*s*：如何可靠地为半监督学习中的自训练选择伪标记数据

Julian Rodemann julian@stat-uni-muenchen.de
Christoph Jansen christoph.jansen@stat.uni-muenchen.de
Georg Schollmeyer georg.schollmeyer@stat.uni-muenchen.de
Thomas Augustin thomas.augustin@stat.uni-muenchen.de
慕尼黑大学统计系，慕尼黑路德维希-马克西米利安大学（*LMU*）

arXiv:2303.01117v1 [stat.ML] 2 Mar 2023

摘要

自训练是半监督学习中一种简单而有效的方法。其思路是通过迭代添加伪标记数据来扩充训练数据。其泛化性能在很大程度上取决于伪标记数据的选择（PLS）。本文旨在提高 PLS 对相关建模假设的稳健性。为此，我们提出选择能够最大化多目标效用函数的伪标记数据。该效用函数旨在考虑不同的不确定性来源，其中三种来源将作更详细的讨论：模型选择、误差累积和协变量偏移。在缺乏关于此类不确定性的二阶信息时，我们还考虑了适用于信念集的广义贝叶斯 α-cut 更新规则这一通用方法。作为概念验证，我们重点展示了三种稳健扩展在模拟数据和真实世界数据上的应用。结果表明，尤其是针对模型选择的稳健性能够显著提升准确率。[1]

关键词： 半监督学习、自训练、广义贝叶斯、模型选择、协变量偏移、广义更新规则

1. 引言

在从图像分类 [69] 、金融计量经济学 [66] 到基因组学 [25] 等众多应用型学习任务中，获取观测数据的标签都十分繁琐。标签数据的稀缺催生了半监督学习 （SSL）这一范式。在 SSL 中，自训练（也称为伪标签）通常被认为是最直接的方法之一 [65, 37, 48]。自训练遵循这样的一般思路：根据模型的预测结果，迭代地为无标签数据分配伪标签。更具体地说，其思想是利用在有标签数据上训练的模型预测无标签数据的类别，并将其中一些预测结果作为伪标记数据纳入训练数据，然后再次对剩余的无标签数据进行预测，这一过程需要一个用于伪标签选择 （PLS）的标准（称为置信度度量），即选择要添加到训练数据中的伪标记实例。大多数此类置信度度量有一个共同点，即它们完全且唯一地源自单个模型。本文旨在针对多个（已拟合的）模型进行伪标记数据的选择，从而使 PLS 对模型不精确性具有稳健性。后者可能有多种来源。第3节详细讨论了如何处理其中的三种来源：模型选择、误差累积和协变量偏移。如果这些来源无法识别，我们将在第4节提出一种针对 PLS 的通用稳健方法，该方法建立在关于信念集和广义贝叶斯推断的丰富文献基础之上。本节剩余部分将讨论相关工作，并在借鉴 [61]的基础上对半监督学习进行形式化介绍。本文最后在第5 和 6章中给出一个简短的真实世界应用，并提出一些总结性意见。

1.1. 半监督学习

绝大多数半监督学习方法都针对分类任务 [73, 13]。我们大致借鉴 [71], , 将半监督学习形式化如下。考虑有标签数据

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n \tag{1}$$

以及无标签数据

$$\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=n+1}^{m} \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^{m-n} \tag{2}$$

来自同一数据生成过程，其中 X 是特征空间，Y 是分类目标空间。半监督学习的目标是利用 D 和 U，学习由 θ 参数化的预测分类函数 $\hat{y}_\theta(x)$。该目标可以分为两方面 [71]。一方面，目标只是为 U 标注（直推式学习）。另一方面，也是更常见的情况，可以同时利用 D 和 U 学习预测函数，以更准确地预测任何未见过的测试

[1] **开放科学：** 所提方法的实现以及用于实验分析的可复现脚本可在以下网址获取：www.github.com/rodemann/reliable-pls。

© J. Rodemann、C. Jansen、G. Schollmeyer 和 T. Augustin。

数据（归纳学习），而不是像经典监督学习那样仅依赖 D。

1.2. 自训练

根据 [53] 和 [73],，半监督学习大致可分为自训练和协同训练。我们将重点讨论前者，其一般思路通常描述为：通过经验风险最小化在 D 上拟合模型，然后利用该模型的预测为 U 添加标签。通常，会根据某种置信度度量，选取 U 中预测置信度最高的实例。预测概率（概率分数）是最常用的度量之一 [71]。此外，也会使用预测方差，以及方差与概率分数的线性组合 [55]。关于将来自 U 的伪标记数据纳入 D， [71] 和 [35] 区分了增量式、批次式和修订机制。增量方法按顺序逐个为实例添加标签，而批次式和修订技术则分别允许一次添加多个数据点或移除数据。此外， [71] 根据标注时使用的已训练分类器数量$\hat{y}$（x），将自训练方法区分为单分类器和多分类器方法。如果使用多个分类器，它们可以基于相同的模型类别，也可以基于多种模型。这称为单模型学习与多模型学习，参见例如 [71]。组合和汇总多个分类器的预测与置信度度量有多种方式。这与我们提出的模型稳健型伪标签选择（PLS）略有关联，参见第 3 和第 4 节。当然，区别在于，我们依据多个模型来选择伪标记数据，而多模型学习则使用多个模型来预测伪标签。

此外，自训练算法可能采用不同的停止准则 [71]。一种简单的做法是为整个集合 U 添加标签并将其纳入训练。另一种做法是，当预测分类器 $\hat{y}(x)$ 不再因 U 而发生变化时停止，并让 U 中剩余的数据保持无标签。在本文中，我们提出了增量式和批次式两种方法，可与任何停止准则配合用于归纳学习。

1.3. 超集学习

超集学习是半监督学习的一种推广。它不使用完全未标注（即完全模糊）的数据 $\mathrm{U} = \{(x_i,\ \mathrm{Y})_i\}_{i=1}^m \in \left(\mathrm{X} \times 2^{\mathcal{Y}}\right)^m$，而是考虑 $\{(x_i,\ Y_i)\}_{i=1}^m \in \left(\mathrm{X} \times 2^{\mathcal{Y}}\right)^m$，其中$Y_i \subseteq \mathrm{Y}$。在这种情况下，$Y_i$ 被视为底层“真实”单元素 y_i的超集，因此得名。超集学习既有乐观变体，也有悲观变体，还有平衡这两种极端情况的方法 [27, 28, 29, 60]。其基本思路是找到超集的单元素表示（通常称为实例化），使其对应于在该表示上训练和评估时预测能力最强（乐观）或最弱（悲观）的模型。在乐观情况下，可以通过最小化经验风险的乐观版本——广义经验风险来实现：$\frac{1}{n}\sum_{i=1}^n L^*(\hat{y}_i,\ Y_i) = \frac{1}{n}\sum_{i=1}^n \min_{y \in Y_i} L(\hat{y}_i,\ y)$，其中 L^*为乐观超集或下确界损失 [9]。

1.4. 鲁棒半监督学习

半监督学习，尤其是自训练的鲁棒性，已得到广泛讨论。 [3] 提出了一种用于伪标签预测的信息论方法，可抵抗协变量偏移。 [74] 致力于通过偏差信息准则进行模型选择，从而降低自训练对建模假设的敏感性。与我们在第 4, [41, 42] 节中对信念集的运用颇为接近的是， 4, [41, 42] 建议将伪标签识别为概率分布集合（“信念自监督学习”）。受一致性正则化 [8, 67, 76], 、超集学习 [27, 28, 29, 60] 和分布对齐 [36], 的启发，“信念自监督学习”旨在降低对单一分布假设的依赖。我们的工作遵循相同的思路，但在概念上有所不同：[41, 42] 首先通过数据增强，利用软标签对训练数据进行模糊化，从而获得集合值预测。在本文中，我们仅在选择阶段利用信念集的表达能力。总体而言，已有大量研究通过贝叶斯技术 [22, 51, 1], 、加权似然 [68], 、条件似然 [23], 和联合混合似然 [2], 提高半监督学习中预测 的鲁棒性。另一方面，关于伪标签的贝叶斯或基于似然的选择 的鲁棒版本，相关研究却很有限（贝叶斯方法）或几乎空白（基于似然的方法），而这正是本文的核心思想。 [40] 的作者利用神经网络的预测分布混合，并采用蒙特卡洛丢弃，在不显式考虑后验预测的情况下模拟贝叶斯设置，从而量化伪标签的不确定性。最近， [52] 提出了 PLS，其依据是伪标签后验预测分布的熵。

[61]从决策理论的角度探讨半监督学习中的伪标签选择（PLS）问题，并提出贝叶斯最优伪标签选择（BPLS）。其思路是通过对参数后验分布进行边缘化，而非仅考虑单个最优参数向量的预测分布，使 PLS 对初始拟合更加稳健。虽然这使我们能够依据给定模型的多种拟合结果来选择伪标记数据，但 BPLS 仍受限于所假设的模型（类型）及其相应的分布假设。这正是本文主体部分将介绍的 BPLS 若干稳健扩展的出发点，即

第 3 节和 4。首先，我们将在第 2.1 节介绍 PLS 的条件视角。这有助于我们将 (B)PLS 理解为决策问题，具体说明见第 2.2 节。

2. 伪标签选择

2.1. 条件伪标签选择

与标准自训练一样，我们首先使用有标签数据 $\mathrm{D} = \{(x_i, y_i)\}_{i=1}^{n}$ 拟合一个参数未知的参数模型 M，其参数向量为 $\theta \in \Theta$。在本文中，我们假设 Θ 是紧致的，并记 $\dim(\Theta) = q$。请注意，贝叶斯推断可以自然地融入这一设定，因为对于给定的参数模型 M，我们可以在 Θ 上指定一个先验函数 $\pi(\theta \mid M)$。我们的目标是通过 θ 学习 $p(\boldsymbol{y} \mid \boldsymbol{x})$ 的条件分布，具体做法是观测 D 中的特征 $\boldsymbol{x} = (x_1, \ldots, x_n)$、$x_i \in \mathrm{X}$，以及类别 $\boldsymbol{y} = (y_1, \ldots, y_n)$、$y_i \in \mathrm{Y}$。正如第 1.1 和 1.2 节所述，我们首先通过有标签数据 D 从 M 估计 $\hat{\theta} \in \Theta$，然后对无标签数据 U 进行预测，并根据某种选择准则，选出我们最有信心的预测数据点（伪标记数据），再将它们添加到 D 中。

最重要的是，在本文中，我们始终不讨论如何利用在 D 上拟合的模型预测 $\mathrm{U} = \{(x_i, \mathrm{Y})i\}_{i=1}^{m}$ 中未知的标签。相反，我们主要关注从已预测的标签中进行选择这一问题。也就是说，我们将 $\mathrm{U} = \{(x_i, \mathrm{Y})i\}_{i=n+1}^{m} \in (\mathrm{X} \times 2^{\mathcal{Y}})^{m-n}$ 中的每个元素与其对应的预测 $\{(x_i, \hat{y})\}i$ 相对应，从而得到 $\hat{}\mathrm{U} = \{(x_i, \hat{y})\}_{i=n+1}^{m} \in (\mathrm{X} \times \mathrm{Y})^{m-n}$。不过，接下来我们仍将使用 U，以强调我们的推理适用于任意函数 $\hat{y}(x)$。这并不意味着我们完全不对预测方法作任何具体说明，参见备注7；在那里，我们将依赖最大似然估计。

2.2. 将 PLS 视为决策问题

遵循 [61], ，我们将伪标签选择形式化为一个具有似然效用的规范决策问题，并以此为若干经典决策准则的稳健扩展奠定基础。

定义1（典范决策问题）将 $(\mathbb{A}, \Theta, u(\cdot))$ 定义为一个决策理论三元组，其中包含行动空间 $\mathbb{A}$、未知的自然状态集合 Θ 以及效用函数 $u\colon \mathbb{A} \times \Theta \to \mathrm{R}$。

在本节中，我们关注的是选择伪标记数据这一决策，其中，一个动作对应于从无标签数据 $\mathbb{A}_{\mathcal{U}} = \{(z, \mathrm{Y}) \mid \exists i \in \{n+1, \ldots\ldots, m\} : (z, \mathrm{Y}) = (x_i, \mathrm{Y})_i \in \mathrm{U}\}$ 中选择一个实例，即将实例作为动作 $\mathbb{A}_{\mathcal{U}} \ni a = (z, \mathrm{Y})$。这与统计决策理论形成鲜明对比，在统计决策理论中，要选择的是估计量而非数据。该决策一个动作的决策由效用函数指导。我们紧密遵循 [61] ，并在一定程度上受 [10, 11]，启发，通过如下方式定义所选数据点 $(z,\mathrm{Y}) = (x_i, \mathrm{Y})_i$的效用：若将其及其预测的伪标签 $\hat{y}(z) = \hat{y}(x_i) = \hat{y}_i \in \mathrm{Y}$ 纳入 $\mathrm{D} \cup (x_i, \hat{y}_i)$ 中，则该数据点与 D 由具有状态（参数） $\theta \in \Theta$ 的模型 M 联合生成的合理性，见定义 2。

定义 2（作为效用的伪标签似然）

给定 $\mathcal{D}$ 和预测泛函 $\hat{y} : \mathcal{X} \to \mathcal{Y}$，我们定义如下效用函数

$$u\colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$$
$$((z, \mathcal{Y}), \theta) \mapsto u((z, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid \theta, M),$$

该函数称为伪标签似然。为便于阐述，下文将用 $\mathit{writel}(i) := p(\mathcal{D} := p(i \mid \theta, M) \cup (x_i, \hat{y}(x_i)) \mid \theta, M)$ 表示伪标签似然。

基于将 PLS 嵌入决策理论这一框架，可以推导出最大-最大或贝叶斯准则等经典决策准则。 [61,第 2 章 2]表明，前者对应乐观超集学习 [28]，后者对应待伪标记数据的后验预测 $p(\mathrm{D} \cup (x_i, \hat{y}_i) \mid \mathrm{D}, M)$，下文将其称为伪后验预测（PPP）。最大-最大准则定义为 Φ_m: $\mathbb{A}_{\mathcal{U}} \to \mathrm{R};\ a \mapsto \max_\theta u(a, \theta)$。$\arg\max \Phi_m$ 中的每个元素称为最大-最大动作。给定 π，贝叶斯准则定义为 Φ_π: $\mathbb{A}_{\mathcal{U}} \to \mathrm{R};\ a \mapsto \mathbb{E}_\pi(u(a, \theta))$。$\arg\max \Phi_\pi$ 中的每个元素称为贝叶斯动作。

3. 稳健伪标签选择：所有可能的似然

在半监督学习（SSL）的常见自训练方法中，或许可以泛化并增强用于预测伪标签的模型。然而，以下我们旨在稳健地选择伪标记数据，参见第2.1节。为此，我们将修改通用效用函数（定义2）及相应的贝叶斯准则 [61, 第2章。2]以考虑三种常见的不确定性和不精确性来源：模型选择、误差累积和协变量偏移。我们不再依赖那些假定“在所有可能情况下”都正确的模型所给出的似然效用，而是建议依赖多个模型给出的所有似然。

3.1. 模型选择：反转奥卡姆剃刀

一个明显且普遍存在的精度不足来源是模型选择：应考虑哪种分布假设下的似然（以及相应的模型）？到目前为止，我们将伪标签似然定义为模型 M 下的似然，我们使用该模型

来预测伪标签。不过，这并非必要。如上所述，我们用于选择伪标记数据的条件方法，使这一选择与伪标签预测完全正交。我们不必将效用函数（见 2）定义为在以下假设下观测到伪标记数据的似然：˜模型 M；我们也可以考虑 M，或多个模型下似然的加权和。接下来，我们先从一般情形入手，考虑任意有限个能够以有意义的方式参数化的不同模型，再逐步讨论嵌套模型，最后讨论嵌套广义线性模型，并探讨如何在 PLS 中考虑它们的设定。

3.1.1. 一般情形

首先考虑任意一个 M_1、……、M_K，$K < \infty$分别定义在相应参数空间上的不同参数模型Θ_1、……、Θ_K。记 $\tilde{\Theta} = \times_{k=1}^{K} \Theta_k$ 它们的笛卡尔˜积，并记 f_k为：$\Theta \to \Theta_k$，$k \in \{1、……、K\}$从笛卡尔积到每个 Θ_k的投影。我们可以很容易地扩展伪标签似然效用（定义 2），以纳入多个模型，从而得到一个多目标决策问题。

定义 3（多模型似然效用）按照定义2，考虑$\mathcal{D}$ 和伪标签ˆ $y \in \mathcal{Y}$，其来自$\hat{y}$：$\mathcal{X} \to \mathcal{Y}$，均为给定。则 K维效用函数

$$
\begin{aligned}
u &: \mathbb{A}_{\mathcal{U}} \times \tilde{\Theta} \to \mathbb{R}^K \\
((x_i, \mathcal{Y})_i, \theta) &\mapsto (\ell(i,1), \ldots, \ell(i,K))'
\end{aligned}
$$

称为多模型似然。为简洁起见，我们将 $l(i,k) = p(i \mid f_k(\theta), M_k) = p(\mathcal{D} \cup (z、\hat{y}(z)) \mid f_k(\theta), M_k) with \theta_k \in \Theta_k$ 表示。再次令 K 表示所考虑的模型数量。

对于定义3 中考虑的这类多目标效用的优化，我们面临的是一个多准则决策问题。针对此类决策问题，有许多求解策略。最近，[32]提出了一种处理多维效用函数的新方法。其思路是：利用所考虑的每个单一维度都具有完全基数性的特点，将效用函数的像嵌入一个偏好系统$\mathcal{A}$，即嵌入一种特定的序理论结构，以便对具有局部基数测量尺度的空间进行建模。[2] 随后，每个此类偏好系统都可由一组函数$\mathcal{N}_{\mathcal{A}}$来描述，其中该集合中的每个元素都具有如下形式$\phi$：$[0,1]^K \to [0,1]$。

因此，最优无标签数据的选择将基于同一集合 $\mathrm{N}_{\mathcal{A}}$。要将前述贝叶斯准则推广到这组效用函数，有许多种可能的做法（相关综述请特别参见 [31]）。这里我们只简要讨论其中一种：它无需作出任何额外假设，并且是将一阶随机占优推广到我们的部分基数设定的一种自然方式。这种推广的思路很直接：如果 π 仍表示自然状态集合 Θ（= 参数）上的先验分布，那么现在，我们不再选择相对于某个固定效用函数使期望效用最大化的无标签数据，而是排除所有被其他数据在所有相容函数 ϕ 下期望占优的无标签数据：$[0,1]^K \to [0,1]$。更正式地说，相对于这一广义随机占优准则，定义 3 中决策问题的解由集合 $\mathbb{A}^{\pi}\mathrm{U}$ 给出，其定义为 $\{a \mid \nexists a' : d_\pi(a', a) \geq 0 \wedge d_\pi(a, a') < 0\}$；其中，对于 $a_1, a_2 \in \mathrm{U}_{\mathbb{A}}$，我们令 $d_\pi(a_1, a_2) = \inf_{\phi \in \mathcal{N}_{\mathcal{A}}} [\mathbb{E}_\pi(\phi \circ u(a_1, \cdot)) - \mathbb{E}_\pi(\phi \circ u(a_2, \cdot))]$。

需要特别注意的是，上述集合中剩余的所有元素相对于所考虑的最优性标准都是不可比较的，也就是说，其中每一个元素都是下一个最佳无标签数据点的同等可信候选。如果领域知识使人们偏好所考虑的某些模型，而这种偏好可以通过权重表示，那么也可以简单地将各个似然标量化如下。

定义4（似然加权和）效用˜函数 u：$\mathbb{A}_{\mathcal{U}} \times \Theta \to \mathrm{R}$；

$$
((x_i, \mathcal{Y})_i, \theta) \mapsto \sum_k w_k \cdot \ell(i,k),
$$

权重为 $w_k \in (0,1)$，$k \in \{1, \ldots, K\}$ 之和为 1 的函数，将被称为似然加权和。

相应的贝叶斯准则（参见第 2.2节）采用多模型似然效用，是伪标记数据的后验预测值的加权和（参见同上）。这一事实可直接由 [61] 中的定理2以及期望值的可加性和齐次性推出。

值得特别注意的是：贝叶斯最优伪标记数据，即根据贝叶斯准则选择伪标记数据这一决策问题的最优解，始终属于此前考虑的集合 $\mathbb{A}^{\pi}\mathcal{u}$。这尤其意味着，上述广义随机占优与基于多模型似然效用的贝叶斯准则是相容的，即后者——无论具体权重如何——都确保不会选择被前者排除的任何标签。这为具体应用中的准则选择提出了如下建议

[2]偏好系统是一个三元组 $\mathrm{A} = [A, R_1, R_2]$，由一个非空集合 $A \neq \emptyset$、其上的预序 $R_1 \subseteq A \times A$ A以及 R_1上的预序 $R_2 \subseteq R_1 \times R_1$ a 组成。直观而言，关系 R_1 刻画了可用的序数信息，而 R_2 则编码了该信息的基数部分。

情形：如果没有基于内容来选择多模型似然效用权重的方法，那么进一步分析应仅依赖集合 $\mathbb{A}^{\pi}\mathcal{U}$。相反，如果可以根据内容确定权重，那么基于多模型似然效用的贝叶斯准则将提供更精确且——此时也不再任意的——结果。现在考虑这样一种情形：通过惩罚模型复杂度来选择权重是自然且恰当的，即嵌套模型。

3.1.2. 嵌套模型

现在设所考虑的模型 M_1、……、M_K、$K<\infty$按如下意义嵌套：$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K$。我们可以对参数空间中由此诱导出的层次结构作如下解释：下方的 $k \in \{1$、……、$K\}$越简单，其假设空间就越简单。为了对 PLS 进行正则化，我们可以对更复杂模型各自的似然效用施加惩罚。在定义 4 中，这可能意味着，例如对所有 $k \in \{1$、……、$K\}$ 设置$w_k = \text{维度}(\Theta_k)$维度$(\Theta_K)$。[3]

然而，我们将采取一种更为稳妥的方法，以保证在所有模型 M_1、……、M_K 下，至少达到预先指定的某一水平 τ 的合理性。为此，在自训练中选择伪标记数据时，我们借鉴了对选择准则进行阈值化的常见做法。也就是说，不仅选择选择函数值最高的一个数据点，还要选择所有超过某一阈值的数据点。我们提出将这一做法扩展为不同模型的似然效用所产生的阈值交集。

定义 5（贝叶斯多模型阈值准则）与定义 *2* 中一样，令 $(x_i,\mathcal{Y})_i$ 为来自 $\mathbb{A}_{\mathcal{U}}$ 的任意决策（选择）。根据定义 *3* 中的多模型似然效用函数，我们为给定 $\mathcal{D}$ 和伪标签 $\hat{y} \in \mathcal{Y}$ 的每个 $(x_i,\mathcal{Y})_i$ 赋予效用。现在考虑以下阈值贝叶斯准则 $\Phi_{\tau,\xi,\pi}$：$\mathbb{A}_{\mathcal{U}} \to \mathrm{R}$；

$$a \mapsto \Phi_{\tau,\xi,\pi}(a) = \begin{cases} 0, & \exists k : \mathbb{E}_{\pi}(\ell(i,k)) < \tau \\ 0.5, & \forall k : \tau < \mathbb{E}_{\pi}(\ell(i,k)) < \xi, \\ 1, & else. \end{cases}$$

再次使用 $l(i,k) = p(i \mid f_k(\theta)$、$M_k)$、$k \in \{1$、……、$K\}$，以及$\xi > \tau$一些预设阈值。

请注意，这分别对应于对所有伪后验预测进行阈值处理，详见第 2.2 节。对于带有 $K = dim(\Theta)$ 的加性回归等参数模型，我们可以利用由参数数量决定的模型层级结构 K。在运行该过程之前（见算法 1），我们首先基于完整模型（含 K 个协变量）对伪标记数据进行阈值处理。我们使用较低的阈值重新拟合如果没有选中任何数据，则使用该阈值。如果选中了正数个数据，则启动算法。我们开始逐步降低k，并在此前所有轮次中选中的伪标记数据均未能通过阈值时终止该过程。算法1 中的伪代码描述了这一过程。

算法 1 逆奥卡姆剃刀

数据： $\mathcal{D}$，$\mathcal{U}$，集合 $S_{K+1} = \emptyset$
结果： $\mathcal{D}$
对于 $k \in \{K,\ldots,1\}$ **执行**
  对于 $i \in \{1,\ldots,|\mathcal{U}|\}$ **执行**
    预测 $\mathcal{Y}$ $\ni \hat{y}_i = \hat{y}(x_i)$
    评估 $\mathbb{E}_{\pi}(l(i,k))$
  end
  选择 $\mathrm{U} \supseteq \mathrm{S}_k = \{(x_i,\hat{y}_i)_i \mid \Phi_{\tau,\xi,\pi}(a) = 1, a \sim i\}$
  如果 $\mathrm{S}_k \cap \mathrm{S}_{k+1} \neq \emptyset$：更新 $\mathrm{D} = \mathrm{D} \cup S_l$
  否则 停止
end

因此，$\bigcap_{k=1,\ldots,K} \mathrm{S}_k \neq \emptyset$我们不仅确保 $\forall k \in \{1,\ldots,K\} : \mathrm{S}_k \neq \emptyset$，还确保。也就是说，在那些能够被某个固定模型同样好地解释的元素中，我们选择那些也能被更简单模型同样好地解释的元素。这可以视为反转奥卡姆剃刀，因为我们关注的是选择数据，而非假设。奥卡姆由来已久的剃刀原则主张，在对单一现象具有相同解释力的相互竞争的假设中，选择假设最少的那个。相反，我们考虑多个现象，并从一组相互竞争的假设中，选择那些仍能由最简单假设解释的现象。奥卡姆剃刀可以通过边际似然（或贝叶斯证据）以贝叶斯统计的方式实现，参见 [33, 54, 44, 43]。回想一下，贝叶斯式伪标记数据选择对应于依据后验预测进行选择，而后验预测不过是伪标记数据似然的边际化版本。[4] 与通过贝叶斯证据或贝叶斯因子（即边际似然之比）进行模型选择一样，我们关注的是数据能在多大程度上由某个模型解释。唯一的区别在于，我们感兴趣的是比较（伪）数据在给定模型时的可能性，而不是反过来。

3.2. 误差累积：在所有后验分布中?

PLS 中最根本的不确定性源于：在后续迭代中，伪标记数据会被视为真实标签。

[3] 在一般的多目标效用情形下，也可以直接对似然 $l(i,1)$、……、$l(i,K)$进行加权，具体见定义 3。

[4] 关于后验进行边缘化。

定义 6（作为效用的多标签似然）与定义 2 中一样，令 (z, Y) 为来自 $\mathbb{A}_{\mathcal{U}}$ 的任意决策（选择）。与定义 2 不同的是，我们现在不仅考虑预测的伪标签 $\hat{y}_i \in Y$，还考虑所有其他假设标签 $\tilde{y}_i \in Y \setminus \{\hat{y}_i\}$。用 $\tilde{y}_{i,j} \in Y$ 表示 $(x_i, Y)_i$ 的所有可能标签，其中 $j \in \{1, \ldots, J\}$ 且 $J = |Y|$。我们通过以下效用函数 u 为每个 $(x_i, Y)_i$ 分配效用：

$$\mathbb{A}_{\mathcal{U}} \times \Theta \to R;$$

$$((z, \mathcal{Y}), \theta) \mapsto \sum_{j=1}^{J} w_j \cdot p(\mathcal{D} \cup (z, \tilde{y}_{i,j}) \mid \theta, M)$$

权重为 $w_j \in (0, 1)$ 且总和为 1。该效用函数称为多标签似然。

同样，相应的贝叶斯准则是伪标记数据的后验预测值的加权和（参见第 2.2 节），这是由于 [61] 中的定理2，以及期望值的可加性和齐次性。权重 $w_j \in (0, 1)$ 的一个合理选择是相应 j个标签的预测概率，即 $p((z, Y) = (z, \tilde{y}_{j,i}))$。这看起来相当直观。然而，尽管可以通过假设分配的标签所具有的可信度来刻画无标签数据点，但仍然必须将它们与实际预测的标签一同加入。

到目前为止，我们放宽了这一限制。值得注意的是，定义 2以及随后所有的论述都依赖于一个模型M以及已经预测的标签。我们放宽了前一种依赖，但保留了后一种依赖。下面的评述对此提出了质疑。

备注7（次优标签并非冗余 ant)

考虑 u: $\mathbb{A}_{\mathcal{U}} \times \Theta \to \mathrm{R}$，根据定义2，令$\hat{y}_i = \hat{y}_{\hat{\theta}_{ML}}(x_i)$，以及$\hat{\theta}_{ML} = \arg\max_\theta p(\mathcal{D} \mid \theta, M)$ 为最大似然估计量。此外，考虑u: $\mathbb{A}_{\mathcal{U}} \times \Theta \to \mathrm{R}$;

$$((z, \mathcal{Y}), \theta) \mapsto u((z, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (z, \tilde{y}(z)) \mid \theta, M),$$

其中 $z = x_i$ 和 $\tilde{y}(z) = \tilde{y}_i = \hat{y}_{\tilde{\theta}}(x_i)$，以及任意次优$\tilde{\theta} \in \Theta$，使得 $p(D \mid \tilde{\theta}, M) \leq p(D \mid \hat{\theta}_{ML}, M)$。相对于 u，最大-最大动作 $a^*_m = max_a\, max_\theta\, u(a, \theta)$通常并不具有低于相对于 $\tilde{u}$的最大-最大动作 $\tilde{a}^*_m = max_a\, max_\theta\, u(a, \theta)$的效用。为说明这一点，令$a^*_m$ 为上述 u 下的最大-最大动作。成立：$a^*_m = max_a\, max_\theta(p(D \cup (z, \hat{y}_i) \mid \theta, M)) = max_a\, p(D \cup (z, \hat{y}_i) \mid \hat{\theta}_{ML}, M)$。类似地，$\tilde{a}^*_m$ 最大化 $p(D \cup (z, \tilde{y}_i) \mid \hat{\theta}_{ML}, M)$。由于在最大似然估计中未考虑 $(z, \tilde{y}_i)$ 和 $(z, \hat{y}_i)$，我们无法对 $u(a^*_m)$与 $\tilde{u}(\tilde{a}^*_m)$之间的关系作出任何判断。贝叶斯准则也同样如此，因为 θ 的后验同样不包含$(z, \tilde{y}_i)$和$(z, \hat{y}_i)$。

受这一观察的启发，我们现在来考虑定义在另一个行动空间上的标准˜效用（定义2） $\mathbb{A}_{\mathcal{U}} =$ $\{(z, y_j) \mid y_j \in \mathrm{Y}$ 和 $\exists i \in \{n+1, \ldots\ldots, m\} : (z, \mathrm{Y}) = (x_i, \mathrm{Y})_i \in \mathrm{U}\}$，以及一种修正后的（完整）贝叶斯准则，该准则考虑了 Y 上的先验 ρ，并按预测步骤中得到的预测分布为标签赋予权重，即 $\mathbb{E}_\rho(\Phi_\pi(a)) = \mathbb{E}_\rho \mathbb{E}_\pi(u(a, \theta))$。

命题8（完全贝叶斯准则等同于加权效用）
在 $w_j = \rho(y_j)$ 的情况下，基于多标签˜效用（定义6）在 $\mathbb{A}_{\mathcal{U}}$ 而非 $\mathbb{A}_{\mathcal{U}}$ 上定义的贝叶斯准则，等同于在 $\mathbb{E}_\rho(\Phi_\pi(a))$ 上的（完全）贝叶斯准则 $\mathbb{A}_{\mathcal{U}}$。

证明 $\mathbb{E}_\rho \mathbb{E}_\pi(u(a, \theta)) = \int \mathrm{Y} \int_\Theta u(a, \theta)\, d\pi(\theta)\, d\rho(y_j) = \int p \cup z\, y_j \mid \theta\, M\, d\rho\, y_j = \sum p \cup z\, y_j \mid \mathrm{Y}($
$M\ \rho\ y_j = \sum_j p \cup z\, y_j \mid M\ \rho\ y_j$ ■
D (,) ,) ( ) Y (D (,)) ( ) (D (,)) ( )

3.3. 协变量偏移

选择准则通常会使某些无标签数据比其他数据更有可能被纳入 [59]。在自训练过程中，这可能导致 X发生分布偏移，通常称为协变量偏移。根据停止准则，这种协变量偏移可能会传递到最终模型，从而可能损害通过可解释机器学习（IML）领域技术获得的模型可解释性。例如，协变量空间中数据稀少的区域会妨碍可靠地估计部分依赖 [18]。值得注意的是，这种分布偏移会影响此前讨论的所有 PLS 选择准则。在本小节中，我们将讨论可能的扩展方法，旨在选择伪标记数据，使其对于以下两者而言均为最优：实际选出的数据 $\mathcal{D}$ 以及一个假设的 独立同分布样本 $\mathcal{D}'$，该样本由随机抽取伪标记数据生成。秉承多模型似然效用（定义3）的思路，并完全类比此前讨论的推广方法，我们可以定义一种多数据似然效用，使 PLS 对协变量偏移具有鲁棒性。上述讨论的决策准则同样适用。另请注意，在这种双目标的特殊情况下，也可以采用区间值效用（损失）函数，例如 [63, 第3节2]。

定义9（多数据似然效用）我们通过以下双目标效用函数，为给定$\mathcal{D}$、$\mathcal{D}'$ 以及预测泛函$\hat{y} : \mathcal{X} \to \mathcal{Y}$ 的每个$(x_i, \mathcal{Y})_i$ 赋予效用 u: $\mathbb{A}_{\mathcal{U}} \times \Theta \to \mathrm{R}^2$;

$$((z, \mathcal{Y}), \theta) \mapsto (\ell_{\mathcal{D}}(i), \ell_{\mathcal{D}'}(i))',$$

与 $l_{\mathcal{D}}(i) = p(\mathrm{D} \cup (x_i, \hat{y}_i) \mid \theta, M)$ 以及 $l_{\mathcal{D}'}(i) = p(\mathrm{D}' \cup (x_i, \hat{y}_i) \mid \theta, M)$。

4. 使用 α-切更新

第 3 节讨论的所有 PLS 稳健扩展，都需要有关所涉及不确定性的某些第二层信息（例如模型选择、先前置信度、协变量偏移）。为了实现一种与具体情形无关且普遍稳健的 PLS 方法，我们转向不精确概率 [75, 5]，以及更具体而言，转向富有成效的先验凸集框架 [38, 39]、[62]、Γ-maximin [64]和 α-cut 更新 [11, 12]。

4.1. 更新 CredalSets

鉴于我们前面已阐述的、对初始模型拟合 $\hat{}\theta$ 普遍持有的怀疑态度，我们希望以一种通用方式减弱似然对后验的影响。这可以通过推广贝叶斯分析 [75, 62, 5] 来实现。同样，我们可以借鉴丰富的决策理论文献，其历史可追溯至 [16, 34, 7]。我们将借鉴最大-期望-最小 [34] 理论，或等价地说，Γ-极大极小理论，例如参见 [64, 7, 19, 72, 26]。为此，我们引入一个先验的凸集 $\Pi \subseteq \{\pi(\theta) \mid \pi(\cdot)$ 是 $(\Theta, \sigma(\Theta))$ 上的概率测度 $\}$，其中 Θ 如上所述是紧致的，而 $\sigma(\cdot)$ 是适当的 σ-代数.[5]。其基本思路如下：观察到数据后，我们依据集合中导致最低后验分布的那个先验来进行选择（采取行动）。换言之，我们针对最坏情况先验进行防范。简而言之，我们选择这样一个伪标记实例：如果我们指定的先验与模型（可能过拟合的）似然相抵触的程度最大，那么该实例就会具有最高的期望效用（似然）。相应的决策准则是 Γ-极大极小准则 $\Phi_\Pi : \mathbb{A}_\mathcal{U} \to \mathrm{R}; a \mapsto \Phi(a) = \mathbb{E}_\Pi(u(a,\theta))$，其中$\mathbb{E}_\Pi(\underline{u}(a,\theta)) = \inf_{\pi \in \Pi} \mathbb{E}(u(a,\theta))$ 表示下期望，以下我们假设它具有仿射超可加性（因此等同于相干下预见）。这将使我们能够利用 [12] 为下预见引入的 α-切更新规则。下期望对应于后验预测，其所依据的后验是通过更新先验 $\pi^*(\cdot) \in \Pi$ 得到的；该先验在最大似然估计量 $\hat{}\theta_{ML}$ 中取值最低。

然而，这种方法可能有些过犹不及，因为其相应的决策可能完全忽略似然，更不用说其对 Π 的高度敏感性了。相反，我们选择一种基于 [11, 12][6] 的信念集更新规则：Cattaneo 的 α 截断更新规则，其中 $\alpha \in (0, 1)$，也称为“软修正” [4]。其基本思路是，仅更新相应边际似然（证据）大于或等于 α 倍相应最大边际似然的那些先验。换言之，在将该集合中的所有先验下期望更新为后验下期望之前，会从下期望集合中剔除其（相对）似然低于 α 的先验。这意味着将所有后验的集合限制为

$$\{\pi \in \Pi \mid m(\pi) \geq \alpha \cdot \max_{\pi} m(\pi)\}, \tag{3}$$

其中 $m(l, \pi) = \int_\Theta l(\theta)\, \pi(\theta)\, d\theta$ 是边际似然。这样一来，我们便能确保不会完全无视似然来做出决策，也就是说，不会基于似然极低的 θ 来做出决策。

此外，α-截更新规则允许动态自适应地选择伪标记数据。请注意，每个预测的伪标签 $\hat{y}$ 都附带一个预测概率 $\hat{}p_{\hat{y}} \in [0,1]$，表示 $\hat{y}$ 是真实标签的概率。选择了具有相应 $(x_i, \mathrm{Y})_i$ 和 $(x_i, \hat{y}_i)$ 的$p_{\hat{y}}$后，该概率代表了我们对数据 $\mathrm{D} \cup (x_i, \hat{y}_i)$ 的信念；后续模型的似然便是基于这些数据指定的。[7] 更一般地，在半监督学习的第 t 次迭代中，我们对伪标记数据的信念为$\prod_{t=1}^{T} \hat{}p_{\hat{y},\, t}$。因此，我们可以在第 t次迭代中，使用 α-截来更新 Π，使得 $\alpha_t = \prod_{t=1}^{T} \hat{}p_{\hat{y},\, t}$。这种自适应 α-截规则的解释如下：我们越不信任伪标记数据，截区间就应该越宽，因为我们不希望仅仅因为可能有缺陷的数据如此表明，就降低某个 θ 的权重。反之，如果我们信任伪标记数据，就可以对截区间采取更严格的限制。尽管这一规则有很强的直观依据，但迄今为止，我们尚未找到这类更新规则的任何保证。因此，接下来我们将基于期望遗憾，推导出一种适用于半监督学习的更新规则，该规则考虑了在 PLS 中来自某个特定模型的特定预测。

4.2. 基于遗憾的更新规则

前面对模型选择（第3.1节）以及次优标签的非冗余性（备注7）的讨论，促使我们修改用于 PLS 的 α-截断更新规则：我们更新 Π，使我们的贝叶斯行动对于遗憾（以比率表示，见定义 10至12）具有某种可量化的保证；该遗憾源于可能错误的标签和可能错误的模型。[8] 因此，我们首先将这两种遗憾量化为定义在 Θ 上的随机变量，然后将总遗憾定义为这两种遗憾的函数的（后验）期望。

[5] Π 中的先验可以反映对先验信息的不确定性，也可以表示近乎无知的先验，参见例如 [6, 46, 45, 56,57, 58]。 [50, 49, 20]已经提出过类似性质的更新规则。值得注意的是， [21, 第 46 页及以下] 将 $\alpha = 1$ 这一特殊情况称为“类型 2 最大似然”，另见 [7, 第 3.5 节4]。

[7] 不妨大胆借鉴频率学派推理，将 $1 - \hat{}p_{\hat{y}}$ 解释为错误频率。[8] 请注意，同时运用两组先验和模型不精确性进行推理，让人想起 [75, 第 8] 章

定义*10*（标签诱导遗憾）考虑 $\tilde{y}_{i,j} \in Y$，它们是 $(x_i, Y)_i$ 的所有可能标签，其中 $j \in \{1, \ldots, J\}$ 且 $J = |Y|$。与备注7中一样，令 $\tilde{u}_j$ 为伪标签似然 $\tilde{u}_j(\cdot, \cdot)$（定义 2），其中 $\tilde{y}_i = \tilde{y}_{i,j}$。此外，设 $\hat{y}_{i,h} = \hat{y}_{\hat{\theta}}(x_i)$ 为实际预测标签，见备注7。对于任意给定的决策 a^* 和任意 θ，函数 $r_l(\cdot, a^*) : \Theta \to R$;

$$\theta \mapsto r_l(\theta, a^*) = \sup_j \frac{u_j(\theta, a^*)}{\tilde{u}_h(\theta, a^*)}$$

称为标签诱导遗憾。

定义*11*（模型诱导遗憾）令 M_1、...、M_K 和 Θ_1、...、Θ_K 分别表示所有正在考虑的模型及其参数空间，以及它们的笛卡尔积 $\tilde{\Theta} = \times_{k=1}^{K} \Theta_k$。与定义*3*中一样，将 $l(i,k) = p(i \mid f_k(\theta), M_K)$ 视为在模型 M_k 下选择 $(x_i, Y)_i$ 且在 Θ_k 上进行投影的似然效用。类比定义*10*，用 M_h 表示实际使用的 ˜ 模型。对于任意决策 $a^* \hat{=} i^*$ 和任意 $\theta \in \Theta$，函数 $r_m(\cdot, a^*) : \Theta \to R$;

$$\theta \mapsto r_m(\theta, a^*) = \sup_k \frac{\ell(i^*, k)}{\ell(i^*, h)}$$

称为模型诱导遗憾。

定义12（半监督学习中的总预测遗憾）将 $\tilde{u}_{j,\,k}(\theta, a^*)$ 表示为在模型 M_k 下，$a^* \hat{=} i^*$ 具有预测 $\tilde{y}_{i^*,\,j}$ 时的效用。函数 $r(\cdot, a^*) : \Theta \to \mathrm{R}$

$$\theta \mapsto r(\theta, a^*) = \frac{\sup_{j,k} \tilde{u}_{j,k}(\theta, a^*)}{\tilde{u}_{h,h}(\theta, a^*)}$$

将称为总（预测）遗憾。

定义13（期望总遗憾泛函）

根据定义*12*，对于给定的 $\Theta \times \Pi \to \mathrm{R}$；$(\theta, \pi) \mapsto \mathbb{E}_\pi(r(\theta, a^*))$ 的期望泛函，其后验为 $a^* \in \mathbb{A}_{\mathcal{U}}$ $\pi \in \Pi$，称为期望总遗憾泛函。

现在，我们可以定义一种 α-切更新规则，使得后验信念集为

$$\Pi_\alpha = \{\pi \in \Pi \mid m(\ell_{h,h}, \pi) \geq \alpha \cdot \sup_{j,k} m(\ell_{j,k}, \pi)\}. \quad (4)$$

请注意，这只是通用α-切更新规则的鲁棒化版本，它遵循方程3，从而使我们能够获得以下关于期望遗憾的保证。

命题14（ α-切的短视遗憾保证）
在上述 α-截断更新规则下，针对任何后验 $\pi \in \Pi$，伪标记数据的贝叶斯最优选择 a^* 的预期总遗憾为 $\mathbb{E}_\pi(r(\theta, a^*)) \leq \frac{1}{\alpha}$。

证明：考虑任意 $\pi \in \Pi_\alpha$。有 $\forall a \in \mathbb{A}_{\mathcal{U}} : m(l_{h,h}, \pi) \geq \alpha \cdot \sup_{j,k} m(l_{j,k}, \pi)$。令 $m(l, \pi)$ 表示关于 π 的边际似然，则有：$\forall a \in \mathbb{A}_{\mathcal{U}} : \int_\Theta l_{h,h}(\theta)\pi(\theta)d\theta \geq \alpha \cdot \sup_{j,k} \int_\Theta l_{j,k}(\theta)\pi(\theta)d\theta \implies \forall a \in \mathbb{A}_{\mathcal{U}}$：$\mathbb{E}_\pi(l_{h,h}(\theta)) \geq \alpha \cdot \mathbb{E}_\pi(\sup_{j,k} l_{j,k}(\theta)) \geq \alpha \cdot \sup_{j,k} \mathbb{E}_\pi(l_{j,k}(\theta))$。特别地，对于 $a^* \in \mathbb{A}_{\mathcal{U}}$，我们有 $\frac{1}{\alpha} \geq \sup_{j,k} \mathbb{E}_\pi(\tilde{u}_{j,k}(a^*, \theta)) \mathbb{E}_\pi(\tilde{u}_{h,h}(a^*, \theta)) \geq \mathbb{E}_\pi(\sup_{j,k} \tilde{u}_{j,k}(a^*, \theta)) \mathbb{E}_\pi(\tilde{u}_{h,h}(a^*, \theta)) = \mathbb{E}_\pi(r(\theta, a^*))$，其中 $l_{j,k}(\theta) = \tilde{u}_{j,k}(a, \theta)$。 ■

α-截更新规则最初被 [12]提出，作为一种连续更新规则。根据 [12,定理 3]，这种连续性也直接适用于基于遗憾的 α-截更新。

4.3. 不精确概率下的广义随机占优

若采用定义 3 中的多模型似然效用（而非其各组成部分的加权和），并结合信念先验 Π，则第 3.1.1 节讨论的广义随机占优准则也很容易作相应调整。在精确 π下使用的解集 $\mathbb{A}^\pi \mathcal{U}$，此处将改为在不精确概率模型下进行稳健化的解集 $\mathbb{A}^\Pi \mathcal{U}$，其定义如下：

$$\{a \mid \nexists a' : D(a', a) \geq 0 \wedge D(a, a') < 0\}, \quad (5)$$

其中，对于 $a_1, a_2 \in \mathrm{U}_{\mathbb{A}}$，我们令 $D(a_1, a_2) = \inf_{\pi \in \Pi} d_\pi(a_1, a_2)$.由 Π 稳健化的集合 $\mathbb{A}^\Pi \mathcal{U}$，其含义与精确 π下集合$\mathbb{A}^\pi \mathcal{U}$ 的含义相似：它包含所有伪标记数据 a；无论先验 $\pi \in \Pi$为何，这些数据都不会被其他伪标记数据 a' 依据广义随机占优严格支配。正式地说，我们因此有 $\mathbb{A}^\Pi \mathcal{U} = \bigcap_{\pi \in \Pi} \mathbb{A}^\pi \mathcal{U}$.与 α-截方法类似，还可以通过将集合 $\mathrm{N}_{\mathcal{A}}$ 和 Π 替换为（合理˜选取的）子集 $\mathrm{N} \subset \mathrm{N}_{\mathcal{A}}$ 和 $\Pi \subset \Pi$，来缩小不可排除的伪标签集合；这些子集用于集合 $\mathbb{A}^\Pi \mathcal{U}$的定义中。

这种缩减集合的做法或许值得采用，因为——具体取决于集合的丰富程度$\mathcal{N}_{\mathcal{A}}$ 和 Π ——集合 $\mathbb{A}^\Pi \mathcal{U}$可能包含过多（甚至可能包含全部）可用选项。对于集合$\mathcal{N}_{\mathcal{A}}$，[31]讨论了一种自然的缩减方法，并在 [30]中作了进一步阐述：这里提出，不考虑底层偏好系统的所有可能表示函数 ϕ，而只考虑那些在底层偏序中，对严格可比性的评估高于某个预设阈值的表示函数 $\xi \in [0, 1]$。对于集合 Π 的缩减，也有一种十分自然的做法：只需将集合 Π˜缩减为方程（ $\Pi = \Pi_\alpha$ 4）中的集合，取 α的某个合理取值即可。当然，也可以结合使用这两种缩减方法。

5. 应用

上述大多数决策准则都需要计算伪后验预测（PPP），其中涉及一个可能难以计算的积分。MCMC采样是贝叶斯方法中通常用来规避此类问题的手段，但这往往也会带来一些计算上的难题。为避免这些问题，我们采用了[61, 章节 3]提出的PPP解析近似。为保证计算可行性，我们进一步用给定 D 的对数似然近似给定 D $\cup(x_i, \hat{y}_i)$ 的对数似然，得到：$p(\mathrm{D} \cup (x_i, \hat{y}_i) \mid \mathrm{D}, M) \approx 2l(\hat{\theta}_{ML}) - \frac{1}{2}\log|I(\hat{\theta}_{ML})|$其中，$I(\hat{\theta}_{ML})$ 为费舍尔信息矩阵。我们使用这一近似来实现上述提出的 PLS 扩展中的三种：多标签效用（定义6），包括未加权和加权求和两种形式（见命题8），以及多模型效用（定义3）。我们将采用这些稳健伪标签选择准则的半监督逻辑回归，与四种常见的 PLS 准则（概率分数、后验预测（贝叶斯行动）、似然（最大-最大行动）和预测方差）以及监督基线进行比较。对于监督基线，我们不进行自训练，仅使用有标签数据训练模型。实验采用模拟的二项分布数据，以及来自 UCI机器学习仓库 [15] 的真实数据集。由于所有数据集中的目标类别都相当均衡，我们比较各方法的（测试）准确率。对于所有数据集，我们通过多轮重复自训练计算测试准确率的平均值，每轮均采用新的随机训练—测试划分。结果令人鼓舞：对于模拟数据，基于多模型伪后验预测的 PLS 准确率最高可提升 15 个百分点。[9]

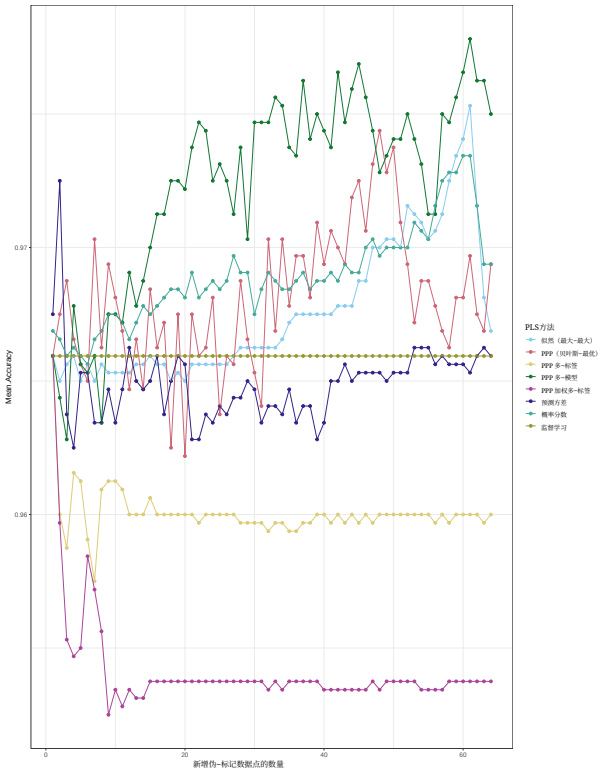


图 1：纸币数据的结果。

这里重点介绍我们的方法在纸币数据上的应用[17, 70] 。该数据包含100张真瑞士法郎钞票和100张伪钞的测量值（对角线长度、底边距和钞票长度）。当前的学习任务是根据这些协变量对钞票进行分类。图1展示了不同PLS方法在80% 无标签数据下的平均准确率（在未见过的测试数据上评估，并对40次重复实验取平均）。对于多模型方法，我们考虑了所有可能的协变量组合。多模型伪后验预测优于其他PLS方法，而多标签扩展甚至不如监督基线。显然，考虑其他方案并不值得

鉴于初始监督准确率如此之高（~ 0.966），分类结果也是如此。

6. 讨论

我们介绍了多种 PLS 的稳健扩展，其中一些也为未来的研究开辟了方向。例如，累积期望误差（第 3.2节）可用于分数贝叶斯更新中的自适应学习率 [77, 24, 14]。未来的工作还可以关注实现和检验基于 α-cuts 的一般性推广，这些推广已在第4节中介绍。总而言之，PLS 似乎是应用现有高效稳健统计学习框架的一个有前景的领域，例如使用信念集进行广义贝叶斯更新，或进行更具体的多模型和多标签稳健化。将 PLS 视为决策问题后，其中大多数方法都有可能轻松迁移到 PLS 中。这不仅可以通过弱化假设、借鉴 Manski 的“可信度递减定律”来提高推断的可信度 [47]。初步证据还表明，这也能够显著提升预测性能。尤其是，我们的实验表明，在 PLS 中考虑替代模型设定以及未被预测的标签，似乎是一种富有前景且卓有成效的方法。此外，还需要进一步研究，以厘清不同类型稳健化之间的相互作用（例如多标签与多模型 PLS 之间的相互作用）。

[9] 关于**所有结果**、实验详情及可复现代码，请参阅 www.github.com/rodemann/reliable-pls。

致谢

Georg Schollmeyer 感谢慕尼黑大学的指导计划所提供的支持。

作者贡献

Julian Rodemann 提出了稳健 PLS 扩展的主要构想，该扩展考虑了模型选择、误差累积和协变量偏移。他起草并撰写了论文的大部分内容。Julian Rodemann 还实现了稳健 PLS，并构思和开展了实验分析。Christoph Jansen 提出了将 α-截断更新规则应用于稳健 PLS 的构想。其基于遗憾的调整由 Julian Rodemann 开发。Christoph Jansen 还撰写了若干段落，介绍如何求解关于广义随机占优的稳健 PLS 问题。Georg Schollmeyer 和 Christoph Jansen 也协助精简了技术记号。Thomas Augustin、Georg Schollmeyer、Christoph Jansen 和 Julian Rodemann 还通过启发性的讨论和细致的校对作出了贡献。

参考文献

[1]Ryan Prescott Adams 和 Zoubin Ghahramani。Archipelago：非参数贝叶斯半监督学习。载于国际机器学习会议，第 1–8 页，2009 年。

[2] Massih-Reza Amini 和 Patrick Gallinari。半监督逻辑回归。载于第*15*届欧洲人工智能会议，第2卷，第11页，2002年。

[3]Gholamali Aminian、Mahed Abroshan、Moham- mad Mahdi Khalili、Laura Toni 和 Miguel Ro- drigues。在协变量偏移下半监督学习的信息论方法。载于国际人工智能与统计会议，第7433–7449页。机器学习研究论文集，2022年。

[4] Thomas Augustin 和 Georg Schollmeyer。评论：关于聚焦、Choquet 容量的软修正与强修正及其在统计学中的作用。2021年。

[5] Thomas Augustin、Frank P. Coolen、Gert de Cooman 和 Matthias C. M. Troffaes 编。不精确概率导论。约翰·威立，奇切斯特，2014年。

[6] Alessio Benavoli 和 Marco Zaffalon。k参数指数族推断中的近似无知先验。《统计学》，49(5):1104–1140，2015年。

[7] James O. Berger。统计决策理论与贝叶斯分析。施普林格，柏林。，第2版，1985年。[8] David Berthelot、Nicholas Carlini、Ian Goodfellow、Nicolas Papernot、Avital Oliver 和 Colin A Raffel。MixMatch：一种半监督学习的整体方法。载于神经信息处理系统进展，第32卷，2019年。[9] Vivien Cabannnes、Alessandro Rudi 和 Francis Bach。通过下确界损失进行部分标注的结构化预测。载于国际机器学习会议，第1230–1239页。机器学习研究论文集，2020年。

[10] Marco EGV Cattaneo。 直接基于似然函数的统计决策。博士论文，苏黎世联邦理工学院，2007年。

[11] Marco EGV Cattaneo。似然决策函数。《电子统计学杂志》，7：2924–2946，2013年。

[12] Marco EGV Cattaneo。不精确概率的连续更新规则。载于知识系统中不确定性信息处理与管理国际会议，第426–435页。施普林格，2014年。

[13] Olivier Chapelle、Bernhard Schölkopf 和 Alexander Zien。半监督学习。自适应计算与机器学习丛书。麻省理工学院出版社，2006年。

[14]R. de Heide、A. Kirichenko、P. Grünwald 和 N.Mehta。安全贝叶斯广义线性回归。载于人工智能与统计学国际会议，第2623–2633页。机器学习研究论文集，2020年。

[15] Dheeru Dua 和 Casey Graff。UCI机器学习仓库，2017年。http://archive.ics.uci.edu/ml。

[16] Daniel Ellsberg。《风险、模糊性与萨维奇公理》。《经济学季刊》，75(4)：643–669，1961年。

[17] Bernhard Flury。《多元统计学：实用方法》。Chapman Hall有限公司，1988年。

[18] JeromeHFriedman。贪婪函数逼近：一种梯度提升机。《统计学年鉴》，29：1189–1232，2001年。

[19] Itzhak Gilboa 和 David Schmeidler。具有非唯一先验的最大最小期望效用。《数理经济学杂志》，18(2)：141–153，1989年。

[20] Itzhak Gilboa 和 David Schmeidler。更新模糊信念。《经济理论杂志》，59(1)：33–49，1993 年。

[21] Irving John Good。《良好思维：概率论及其应用基础》。明尼苏达大学出版社，1983 年。

[22] Jonathan Gordon 和 José MiguelHernández-Lobato。结合深度生成模型和判别模型进行贝叶斯半监督学习。《模式识别》，100:107156，2020年。

[23] Yves Grandvalet 和 YoshuaBengio。通过熵最小化进行半监督学习。载于神经信息处理系统进展，第17卷，2004年。

[24] P. Grünwald。安全贝叶斯方法。载于国际算法学习理论会议，第169–183页。施普林格，2012年。

[25] Hüseyin Anil Gündüz、Martin Binder、Xiao-Yin To、René Mreches、Philipp C Münch、Alice C McHardy、Bernd Bischl 和 Mina Rezaei。Self-genomenet：利用反向互补上下文预测对核苷酸级基因组学数据进行自监督学习。2021年。[26] Peijun Guo 和 Hideo Tanaka。基于区间概率的决策。欧洲运筹学杂志，203(2)：444–454，2010年。[27] Eyke Hüllermeier。从不精确和模糊观测中学习：通过广义损失最小化进行数据消歧。《国际近似推理杂志》，55：1519–1534，2014年。[28] Eyke Hüllermeier 和 Weiwei Cheng。基于广义损失最小化的超集学习。载于欧洲机器学习与数据库知识发现联合会议，第260–275页。施普林格，2015年。[29] Eyke Hüllermeier、Sébastien Destercke 和 Ines Couso。从不精确数据中学习：对乐观和悲观变体的调整。载于可扩展不确定性管理国际会议（*SUM*），第266–279页。施普林格，2019年。[30] C. Jansen、H. Blocher、T. Augustin 和 G. Schollmeyer。复杂结构化偏好的信息高效学习：征询程序及其在不确定性下决策中的应用。《国际近似推理杂志》，144：69–91，2022年。[31] Christoph Jansen、Georg Schollmeyer 和 Thomas Augustin。具有部分序数偏好和部分基数偏好的严重不确定性下决策概念。《国际近似推理杂志》，98：112–131，2018年。

[32] Christoph Jansen、Georg Schollmeyer 和 Thomas Augustin。严重不确定性条件下的多目标决策。*arXiv* 预印本，2022年。

[33] HaroldJeffreys。《概率论》。克拉伦登出版社，牛津，1939年。

[34] Eduard Kofler 和 Günter Menges。不完全信息下的决策。施普林格，1976年。

[35] Georgios Kostopoulos、Stamatis Karlos、Sotiris Kot- siantis 和 Omiros Ragos。半监督回归：近期综述。《智能与模糊系统》，35(2):1483–1500，2018年。

[36] Alex Kurakin、Colin Raffel、David Berthelot、Ekin Do- gus Cubuk、Han Zhang、Kihyuk Sohn 和 Nicholas Carlini。RemixMatch：结合分布匹配与增强锚定的半监督学习。载于国际表征学习会议（*ICLR*），2020年。[37] Dong-Hyun Lee。伪标签：一种简单高效的深度神经网络半监督学习方法。载于国际机器学习会议表征学习挑战研讨会，第2卷，第896页，2013年。[38] Isaac Levi。论不确定概率。《哲学杂志》，71：391–418，1974年。[39] Isaac Levi。《知识事业：关于知识、信念概率与机会的论文》。麻省理工学院出版社，1980年。[40] Shuangshuang Li、Zhihui Wei、Jun Zhang 和 Liang Xiao。深度半监督学习的伪标签选择。载于*IEEE*信息学与计算进展国际会议（*PIC*），第1–5页。IEEE，2020年。[41] Julian Lienen 和 Eyke Hüllermeier。信念自监督学习。载于神经信息处理系统进展，第14370–14382页，2021年。[42] Julian Lienen、Caglar Demir 和 Eyke Hüllermeier。保形信念自监督学习。*arXiv*预印本 *arXiv:2205.15239*，2022年。[43] Sanae Lotfi、Pavel Izmailov、Gregory Benton、Micah Goldblum 和 Andrew Gordon Wilson。贝叶斯模型选择、边际似然与泛化。载于国际机器学习会议，第14223–14247页，2022年。[44] David JC MacKay、David JC Mac Kay 等。信息论、推断与学习算法。剑桥大学出版社，2003年。

[45] Francesca Mangili。一种接近无知先验的非参数回归高斯过程模型。载于*2015*年国际不精确概率研讨会（*ISIPTA*）。

[46] Francesca Mangili 和 Alessio Benavoli。单纯形上的新型近乎无知先验模型。国际近似推理杂志，56：278–306，2015年。

[47] Charles Manski。概率分布的部分识别。施普林格，2003年。

[48] David McClosky、Eugene Charniak 和 Mark J ohnson。用于句法分析的有效自训练。载于北美计算语言学协会人类语言技术会议主会场论文集，第152–159页，2006年。

[49] Serafin Moral。根据条件概率凸集计算不确定性区间。载于人工智能中的不确定性，第199–206页。爱思唯尔，1992年。[50] Serafín Moral 和 Luis M De Campos。更新不确定信息。载于基于知识系统中不确定性的信息处理与管理国际会议，第58–67页。施普林格，1990年。[51] Yin Cheng Ng、Nicolò Colombo 和 Ricardo Silva。基于图高斯过程的贝叶斯半监督学习。载于神经信息处理系统进展，第31卷，2018年。[52] Gaurav Patel、Jan P Allebach 和 Qiang Qiu。Seq-ups：面向半监督文本识别的序贯不确定性感知伪标签选择。载于*IEEE/CVF* 冬季计算机视觉应用会议，第6180–6190页，2023年。[53] Nitin Namdeo Pise 和 Parag Kulkarni。半监督学习方法综述。载于计算智能与安全国际会议，第2卷，第30–34页。IEEE，2008年。[54] Carl Rasmussen 和 Zoubin Ghahramani。奥卡姆剃刀。神经信息处理系统进展，第13卷，2000年。[55] Mamshad Nayeem Rizve、Kevin Duarte、Yogesh S Rawat 和 Mubarak Shah。为伪标签辩护：面向半监督学习的不确定性感知伪标签选择框架。载于国际学习表征会议（*ICLR*），2020年。

[56] Julian Rodemann。随机无导数优化的鲁棒推广。硕士论文，慕尼黑路德维希-马克西米利安大学，2021年。

[57] Julian Rodemann 和 Thomas Augustin。贝叶斯优化中模型设定不精确性的处理。在不精确概率国际研讨会（*ISIPTA*）上发表的海报，2021年。

[58] Julian Rodemann 和 Thomas Augustin。贝叶斯优化中高斯过程不精确性的处理。载于知识建模与决策中综合不确定性国际研讨会（*IUKM*），第92–104页。施普林格，2022年。

[59] Julian Rodemann、Sebastian Fischer、Lennart Schnei- der、Malte Nalenz 和 Thomas Augustin。《并非所有数据生而平等：抽样理论对自适应机器学习的启示》。发表于国际统计与数据科学会议（*ICSDS*）的海报，数理统计学会（*IMS*），2022年。[60] Julian Rodemann、Dominik Kreiss、Eyke Hüllermeier 和 Thomas Augustin。通过谨慎超集学习进行逐层数据消歧。载于国际可扩展不确定性管理会议（*SUM*），第263–276页。施普林格，2022年。[61] Julian Rodemann、Jann Goschenhofer、Emilio Dori- gatti、Thomas Nagler 和 Thomas Augustin。贝叶斯PLS！近似贝叶斯最优伪标签选择（PLS）。*arXiv*预印本，2023年。[62] Fabrizio Ruggeri、David Ríos Insua 和 Jacinto Martín。稳健贝叶斯分析。《统计学手册》，25：623–667，2005年。[63] Patrick Michael Schwaferts 和 Thomas Augustin。基于不精确假设的贝叶斯决策与简单假设。载于不精确概率国际研讨会：理论与应用，第338–345页。机器学习研究论文集，2019年。[64] Teddy Seidenfeld。两种决策规则之间的对比：用于（凸）概率集的 γ-极小极大与 e-可采纳性。*Synthese*，140(1/2)：69–88，2004年。[65] Weiwei Shi、Yihong Gong、Chris Ding、Zhiheng MaX- iaoyu Tao 和 Nanning Zheng。使用极大极小特征进行传导式半监督深度学习。载于欧洲计算机视觉会议，第299–315页，2018年。

[66] Hyunjung Shin、Tianya Hou、Kanghee Park、Chan- Kyoo Park 和 Sunghee Choi。基于半监督学习预测原油价格的变动

方向。决策支持系统，55(1):348–358，2013年。

[67]Kihyuk Sohn、David Berthelot、Nicholas Carlini、Zizhao Zhang、Han Zhang、Colin A Raffel、Ekin Do- gus Cubuk、Alexey Kurakin 和 Chun-Liang Li。Fix-match：通过一致性和置信度简化半监督学习。载于神经信息处理系统进展，第 33 卷，第 596–608 页，2020年。

[68] Nataliya Sokolovska、Olivier Cappé 和 François Yvon。判别式概率模型中半监督学习的渐近性质。载于国际机器学习会议，第984–991页，2008年。[69] Farzin Soleymani、Mohammad Eslami、Tobias Elze、Bernd Bischl 和 Mina Rezaei。用于大规模医学图像自标注的深度变分聚类框架。载于医学成像*2022*：图像处理，第12032卷，第68–76页。SPIE，2022年。[70] Cristina Tortora、Ryan P Browne、Brian C Franczak、Paul D McNicholas、维护者 Cristina Tortora，以及依赖项 Bessel。《mixghd》软件包。2014年。[71] Isaac Triguero、Salvador García 和 Francisco Herrera。半监督学习的自标注技术：分类体系、软件与实证研究。知识与信息系统，42(2)：245–284，2015年。[72] Matthias C.M. Troffaes。利用不精确概率进行不确定性下的决策。《国际近似推理杂志》，45(1)：17–29，2007年。[73] Jesper E Van Engelen 和 Holger H Hoos。半监督学习综述。机器学习，109 (2)：373–440，2020年。[74] Vincent Vandewalle、Christophe Biernacki、Gilles Celeux 和 Gérard Govaert。用于在半监督分类中选择生成模型的预测偏差准则。计算统计与数据分析，64：220–236，2013年。[75] Peter Walley。非精确概率下的统计推理。Chapman & Hall，1991年。[76] Bowen Zhang、Yidong Wang、Wenxin Hou、HAO WU、Jindong Wang、Manabu Okumura 和 Takahiro Shinozaki。Flexmatch：通过课程式伪标注提升半监督学习。载于神经信息处理系统进展，第18408–18419页，2021年。

[77] T. Zhang。从 ε-熵到 KL-熵：最小信息复杂度密度估计分析。统计年鉴，34(5)：2180–2210，2006年。